Identification de scripteurs basée sur une distribution probabiliste de prototypes d'allographes

G. X. Tan, C. Viard-Gaudin, A. C. Kot

Identification de Scripteurs basée sur une Distribution Probabiliste de Prototypes d'Allographes

Guo Xian Tan[1,2] – Christian Viard-Gaudin[2] – Alex C. Kot[1]

[1] Nanyang Technological University of Singapore

[2] IRCCyN, UMR CNRS 6597, Ecole Polytechnique de l'Université de Nantes

tanguoxian@pmail.ntu.edu.sg, christian.viard-gaudin@univ-nantes.fr, eackot@ntu.edu.sg

Résumé : *La diversité des solutions de saisie ouvre la voie à l'acquisition et au stockage d'une quantité sans cesse croissante de documents de différentes natures contenant de l'écriture manuscrites en-ligne. Face à ce flux de documents, il est nécessaire de proposer des fonctionnalités permettant un accès intelligent aux bases de données où ils sont entreposés. L'une de ces fonctionnalités consiste à retrouver les documents provenant de la main d'un scripteur donné. Ce papier propose une méthode permettant d'identifier le scripteur d'un texte quelconque de quelques lignes en le comparant à des écritures de références. La comparaison est basée sur une mesure de mise en correspondance des distributions des allographes de lettres représentatifs des styles d'écriture. Pour cela, chaque échantillon de lettre est classé de manière probabiliste parmi les prototypes disponibles pour cette lettre. Le système proposé atteint un taux d'identification de 99,2 % sur une base de référence de 120 scripteurs.*

Mots-clés *:* identification de scripteur, recherche d'information, écriture manuscrite en-ligne, k-plus-proches-voisins, allographe.

1 Introduction

Les progrès technologiques réalisés dans le domaine des environnements de saisie, favorisant des interactions transparentes, mobiles et sans discontinuité, repositionnent l'écriture manuscrite comme une modalité très pertinente pour de nombreuses applications, tant professionnelles que plus personnelles [Oviatt 00]. Ainsi, après l'émergence des assistants personnels puis des Tablettes PC au début des années 2000, on assiste aujourd'hui à un fort développement des téléphones portables disposant d'interface de saisie tactile. Une autre technologie est en fort développement, c'est celle des solutions de saisie de type papier+stylo digital qui ouvrent sur des applications à très large échelle et pour des documents de complexité quelconque [Jain 03]. L'ergonomie de ces solutions se rapproche maintenant de l'usage d'un stylo conventionnel. Deux technologies complémentaires sont à mentionner, celles dépendant d'un papier assurant le repérage spatial et celles déportant ce repérage dans un dispositif annexe intégrant des capteurs ultra-sonores.

L'augmentation constante des productions manuscrites issues de ces dispositifs nécessite des outils de gestion efficaces pour la bonne indexation et la recherche dans les bases de données. De nombreuses fonctionnalités sont souhaitables. On peut évoquer la catégorisation des documents, la recherche par mots-clés, mais également la capacité à retrouver les documents rédigés par une personne donnée. Cette problématique a jusqu'à présent été abordée essentiellement dans le domaine de l'écriture hors-ligne, nous proposons ici de la développer dans le cadre de signaux en-lignes.

L'accès à l'identité du scripteur ayant composé un document apporte de la valeur ajoutée au document. Tout d'abord, du point de vue de la sécurité de l'information, l'identification de scripteurs s'inscrit dans une politique de gestion des droits numériques et de prévention de la fraude et du vol d'identité. Par exemple, l'une des applications pourrait être de traiter l'identité des étudiants composant à un examen à des fins de contrôle. Deuxièmement, dans des environnements où de grandes quantités de documents, formulaires, notes et procès-verbaux de réunions sont constamment en cours de traitement et de gestion, connaître l'identité du rédacteur donnerait une valeur supplémentaire pour distribuer de façon adaptée ces documents ou remonter à la source d'une information.

Le reste de ce document est organisé comme suit : la section 2 rappelle les principaux travaux sur ce sujet, ensuite la méthodologie proposée est détaillée en section 3, tandis que les résultats expérimentaux sont présentés dans la section 4. Enfin, les discussions et les pistes à explorer sont données dans la section 5.

2 Les méthodes existantes

A un premier niveau, on distingue deux types de systèmes d'identification/vérification de scripteurs : ceux qui sont indépendants du texte écrit et ceux qui reposent sur un texte imposé. Les signatures sont un cas

particulier de cette seconde catégorie. A l'inverse, le système que nous proposons ici s'inscrit dans la première catégorie. Un second niveau de différenciation concerne la nature hors-ligne ou en-ligne du signal d'écriture. De nombreux systèmes ont été proposés pour traiter des images de documents [Hochberg 97], [Busch 05], [Bulacu 07], [Niels 07], beaucoup moins de travaux concernent l'écriture en-ligne [Jain 03]. Le système proposé ici s'intéresse aux documents en-ligne, il s'agit d'une extension des travaux de [Chan 08], la taille de la base de référence a été élargie et le calcul du vecteur caractéristique du style du scripteur utilise une méthode plus précise. Enfin, les différents systèmes se distinguent par le type d'approches mises en œuvre. Grossièrement, il en existe deux types. Certaines vont extraire des caractéristiques globales, significatives d'un point de vue macroscopique [Pitak 04], [Yasushi 03]. Il s'agit par exemple de la densité des lignes, de la fluctuation des lignes de bases, du respect de l'indentation. Des mesures d'entropie, de paramètres de textures [He 08] peuvent contribuer à ce type de description. A l'inverse, il peut être intéressant d'avoir une vue beaucoup plus locale et de baser l'identification sur des caractéristiques apparaissant à une échelle microscopique. C'est le cas des approches se basant sur les singularités des graphèmes composant l'écriture. Là encore le niveau de granularité peut être variable. Il s'agit la plupart du temps de graphèmes extraits par des algorithmes de segmentation simples [Bensefia 05], où alors manuellement [Schomaker 04]. L'originalité de notre contribution réside dans le choix d'extraire des lettres grâce à un processus de segmentation/étiquetage automatique du texte en caractères. Nous pensons en effet que le niveau lettre porte une information biométrique de nature très stable.

3 La Méthode Proposée

La méthode d'identification de scripteurs proposée repose sur trois étapes, cf. FIG. 1 . La première consiste à sélectionner lettre par lettre les prototypes caractéristiques des allographes de cette lettre. La seconde étape représente le codage sur la base des allographes-prototypes des documents de référence et de test. Enfin, la troisième étape correspond à la classification du document de test vis-à-vis des documents de référence.

Pour construire les prototypes définissant les allographes de lettres, nous avons utilisé les mots de la base IRONOFF [Viard-Gaudin 99]. Cette base comporte 16 585 mots écrits par 373 sujets. Chaque mot dont on connaît le vrai label est ensuite segmenté automatiquement en lettres. Nous disposons alors de 89 760 caractères répartis sur les différentes lettres de l'alphabet. La lettre la plus fréquente est bien entendue la lettre *'e'* avec 12 161 occurrences tandis que la moins fréquente est le *'w'* avec seulement 139 échantillons. Sur chacun des sous-ensembles correspondant aux 26 lettres minuscules de l'alphabet, un algorithme de *clustering* non-supervisé est mis en œuvre. C'est lui qui permet de définir les N prototypes d'allographes qui permettront de représenter les styles d'écritures de chaque lettre. Ces prototypes sont communs à l'ensemble des scripteurs. L'usage spécifique que fera chaque scripteur de ces prototypes génériques permettra l'identification de celui-ci. Dans les expériences relatées dans cet article, nous avons utilisé un algorithme de type K-moyennes en faisant varier le nombre de *clusters* (de prototypes d'allographes) entre 2 et 60 par lettre. La distance utilisée est la distance euclidienne dans l'espace de description des caractères. Chaque caractère est représenté par un vecteur de 210 composantes, après avoir normalisée et ré-échantillonnée la trajectoire sur 30 points et extrait 7 caractéristiques par point. Les caractéristiques utilisées sont les coordonnées x et y, la direction ($\cos\theta$ et $\sin\theta$) de la tangente, la courbure ($\cos\Delta\theta$ et $\sin\Delta\theta$) et l'état posé/levé du stylo en ce point [Chan 08].

Concernant la seconde étape, chaque document, qu'il soit issu de la base des documents de référence, ou qu'il soit un document de test dont on veut identifier le scripteur va être projeté dans l'espace des prototypes d'allographes. En ne considérant que les lettres minuscules, cet espace est de dimension $26 \times N$. C'est dans cet espace que se fera la mise en correspondance entre un document de test et les documents de référence pour permettre d'ordonner ceux-ci vis-à-vis du document de test. La figure 1 montre le schéma de l'ensemble de ce système d'identification du scripteur.

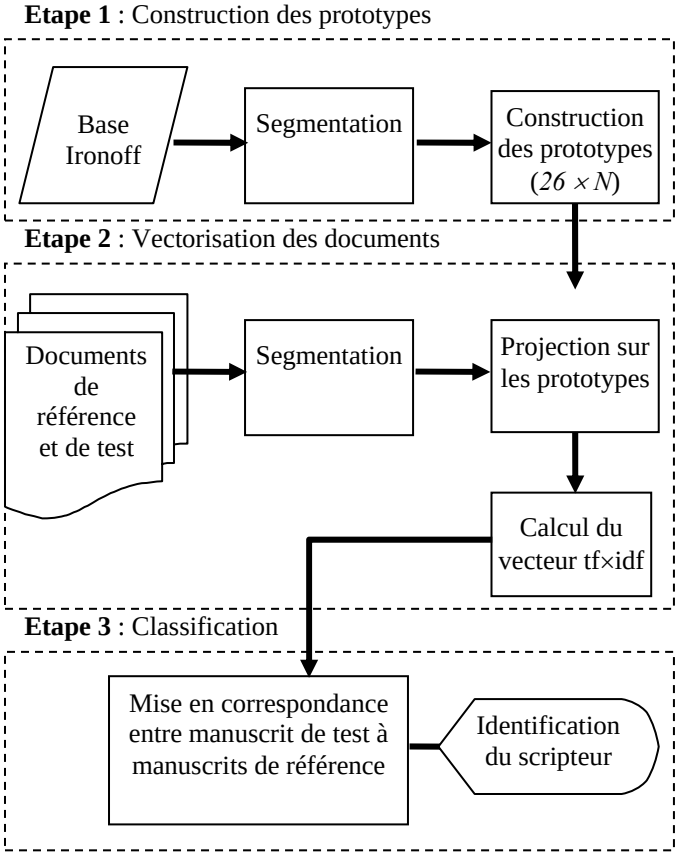


FIG. 2 – Schéma de la méthodologie proposée

Un des points importants de la méthode réside dans la capacité à reconnaître, segmenter et étiqueter automatiquement le texte au niveau caractère. Toutes ces tâches sont confiées à un moteur de reconnaissance industriel (MyScript Builder) [Vision 07].

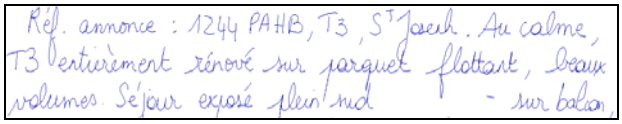


FIG. 3 – Exemple de texte

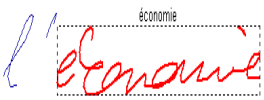
La Fig. 2 montre un exemple de texte dont on veut identifier le scripteur. La Fig. 3 affiche le résultat de la reconnaissance au niveau mot, tandis que la Fig. 4 présente deux des caractères issus de la segmentation.

FIG. 4 – Exemple de résultat de reconnaissance au niveau mot

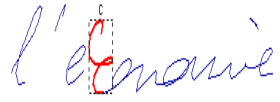
FIG. 5 – Segmentation des caractères, cas du *'f'* et du *'o'*

Bien entendu, l'outil de reconnaissance ne réalise pas une segmentation et une reconnaissance parfaite. Il en résulte des erreurs dans l'affectation des séquences de points (segments) à la lettre qui aurait dû lui correspondre. Par exemple sur la Fig. 5, on montre l'effet d'une erreur de segmentation : l'accent de la lettre *'é'* du mot « *économie* » qui a été correctement reconnu est affecté à tort à la lettre suivante *'c'*. Les performances de l'outil de reconnaissance sont évaluées dans la section 4.

a) texte à reconnaître



b) mot reconnu



c) segmentation incorrecte

FIG. 6 – Erreur de segmentation

Une fois que tous les caractères d'un document sont extraits, nous projetons chacun de ces caractères dans la base des prototypes de la lettre qu'il représente, et nous mesurons une similitude entre ce caractère et chacun des prototypes de cette lettre. Cette mesure est normalisée pour être interprétée comme la probabilité que le prototype ait émis ce caractère [Han 06], [Hoppner 99].

$$p(p_k|x) = \frac{\exp(-\beta \times dist(x, p_k))}{\sum_{k'=1}^{N} \exp(-\beta \times dist(x, p_{k'}))} \quad (1)$$

Dans cette formule x représente un échantillon d'une lettre dont on veut calculer la probabilité qu'il ait été généré par le prototype p_k. N est le nombre de prototypes d'allographes pour cette lettre. La distance $dist(x,p_k)$ est la distance de Mahalanobis entre le point x et le cluster de prototype p_k, elle est calculée dans l'espace de représentation des caractères, ici de dimension *210*. Dans l'équation (1), $\beta > 0$ est un paramètre de réglage fixant la sélectivité des fonctions exponentielles. Ainsi, avec une faible valeur de β, la masse de probabilité sera distribuée sur un grand nombre de prototypes, tandis qu'avec une valeur de β plus élevée les prototypes les plus proches seront davantage concernés. Nous avons réglé expérimentalement $\beta = 0{,}01$ dans nos expériences.

En procédant ainsi, on ne prend pas une décision stricte en ne considérant que le plus proche voisin, comme cela avait été fait dans [Chan 08], mais tous les prototypes sont partiellement considérés.

On peut alors obtenir la fréquence du prototype p_k, soit tf_k (term frequency) sur l'ensemble du document en sommant sur tous les échantillons appartenant à la même lettre.

$$tf_k = \frac{1}{M} \sum_{x}^{M} p(p_k|x) \quad (2)$$

M est le nombre de caractères du document correspondant à la même lettre de l'alphabet.

Enfin, ces termes sont pondérés par un facteur idf_k (inverse document frequency) indiquant le caractère fréquent ou non de ce terme dans l'ensemble de la base des documents de référence. Il en résulte un vecteur de dimension $26 \times N$: $[tf_{k,\alpha} \times idf_{k,\alpha}]$, avec $k \in [1, N]$, et $\alpha \in$ ['a', 'z'], décrivant le style de l'écriture du document, en considérant les 26 lettres minuscules 'a' à 'z'. L'utilisation de descripteurs de type *tf×idf* est classique en recherche d'informations [Salton 88].

A partir de cette représentation vectorielle, il est possible de calculer la distance entre le document de test, t, et chaque document de référence d. Le scripteur dont le document minimise la distance $dist(t, d)$ sera considéré en première position. Nous avons envisagé trois métriques pour calculer cette distance. D'une part classiquement la distance euclidienne, par ailleurs comme les composantes de ce vecteur proviennent de distribution, nous avons aussi utilisé la distance du Chi^2, et enfin pour la même raison la pseudo-distance de Kullback-Leibler [Cover 91]. Le calcul de ces distances s'effectue en ne prenant en compte que les lettres de l'alphabet effectivement représentées à la fois dans le document de test, t, et le document de référence d [Chan 08].

4 Résultats expérimentaux

Afin d'évaluer les performances de ce système d'identification du scripteur, nous utilisons une base de documents issus de 120 scripteurs. Chacun de ces scripteurs a rédigé 2 documents, sur des sujets libres, totalement distincts et à 2 moments différents. La taille des documents est variable, elle s'échelonne entre 86 caractères pour le plus court à 972 caractères pour le plus long. Ces documents ont été collectés avec une technologie de type papier et stylo digitaux, essentiellement dans le même milieu culturel, celui des étudiants et des enseignants. On peut donc s'attendre à une certaine homogénéité si ce n'est dans le type d'écriture au moins dans l'usage qui est fait de l'écriture manuscrite, rendant ainsi le problème de l'identification non trivial. Les exemples présentés aux Fig. 2 et Fig. 5 sont issus de ces documents. Deux sous-bases ont été constituées, chacune de 120 documents, contenant un document provenant de chaque scripteur. L'une de ces sous-bases correspond à la base de référence, l'autre servira de base de test pour évaluer la capacité du système à retrouver dans la base de référence le bon scripteur.

Sur l'ensemble de ces deux bases, nous avons d'abord calculé le taux de reconnaissance de l'outil de qui nous sert à faire la segmentation automatique en lettres. Nous l'utilisons pour cela dans sa version standard avec les ressource linguistiques de base qui ne sont pas spécifiquement adaptées aux textes que nous traitons ici. Le taux de reconnaissance au niveau lettre est alors de *91.0 %*. Les lettres mal reconnues (9 %) vont bruiter les assignations dans les bases lettres correspondantes.

Nous comparons maintenant les résultats en termes d'identification du scripteur entre la version de base (1-PPV) et la version proposée ici où chaque échantillon de lettre est distribué sur les différents prototypes de cette lettre en fonction de sa similitude avec eux (eq. 1), TAB. 1.

1-PPV[1] Dist. euclid.	Distribution sur l'ensemble des prototypes		
	Distance euclidienne	Divergence KL	Distance Chi^2
96.7 %	98.3 %	91.7%	99.2 %
116/120	118/120	110/120	119/120

TAB. 1 – Taux d'identification en première position

Avec la méthode ne considérant que le seul plus proche prototype pour décrire les allographes d'un scripteur, le taux d'identification est de 96.7 %. Cela correspond à 4 scripteurs sur les 120 de la base de test qui ne sont pas retrouvés en première position parmi les 120 documents de la base de référence. Le taux d'erreur est divisé par deux avec la méthode proposée ici, puisque seuls deux scripteurs échappent à la première position lorsque l'on utilise la distance euclidienne. Avec la distance du Chi^2, on diminue encore l'erreur : un seul scripteur n'est pas retrouvé en première position. De plus, le vrai scripteur correspondant est positionné en seconde position de la liste des documents de référence. Par contre, les résultats sont sensiblement moins bons lorsque la divergence de Kullback-Leibler est utilisée, il est possible que la non-symétrie de cette métrique soit un facteur défavorable.

L'amélioration que nous constatons par rapport à la méthode originale est à mettre au crédit d'une meilleure représentation dans l'espace des prototypes. Initialement cet espace était quantifié de façon discrète : un symbole discret représentait tout un sous-espace des caractéristiques (dim : 210), il en résultait un changement brutal de représentant lors de petits déplacements dans cet espace. Nous sommes passés avec la méthode proposée à une représentation continue dans cet espace en estimant la probabilité de chaque style vis-à-vis de l'échantillon de caractère étudié. De cette façon lorsqu'un style d'écriture se trouve en frontière de plusieurs prototypes, la stabilité de la description est bien meilleure.

4.1 Influence du nombre de prototypes

Dans toutes les expériences précédentes, à chaque lettre de l'alphabet est assigné un ensemble de $N = 10$ prototypes. Ce choix résulte d'une étude expérimentale dont les résultats sont reportés à la FIG. 6. Dans cette étude, le nombre de prototypes par lettre évolue de 2 à 60. Afin de vérifier la stabilité vis-à-vis de la base de test, nous avons subdivisé aléatoirement celle-ci en deux parties égales de 60 scripteurs tout en conservant les 120 documents dans la base de référence. On observe sensiblement le même comportement sur les 2 sous-bases, cf. FIG. 6, à savoir qu'avec moins de 10 prototypes, les performances sont dégradées, et de même au delà de 30. Ainsi, avec trop peu de prototypes la précision de la représentation est trop grossière, le biais de l'estimateur des fonctions de densité de probabilité (*ddp*) entraîne ces piètres performances. A l'inverse lorsque le nombre de prototypes est trop élevé, l'estimateur est trop sensible, le bruit présent dans les données se trouve trop pris en compte. Le nombre de 10 prototypes apparaît réaliser le meilleur compromis entre biais et variance dans l'estimation des *ddp*. On pourrait pousser plus loin l'analyse sur le nombre de prototypes en travaillant non plus globalement sur l'alphabet, mais lettre par lettre. Certaines lettres, plus complexes que d'autres, présentent davantage de variations allographiques. C'est sans doute le cas par exemple de la lettre *'f'* vis-à-vis de la lettre *'e'*.

[1] 1-Plus Proche Voisin [Chan 08]

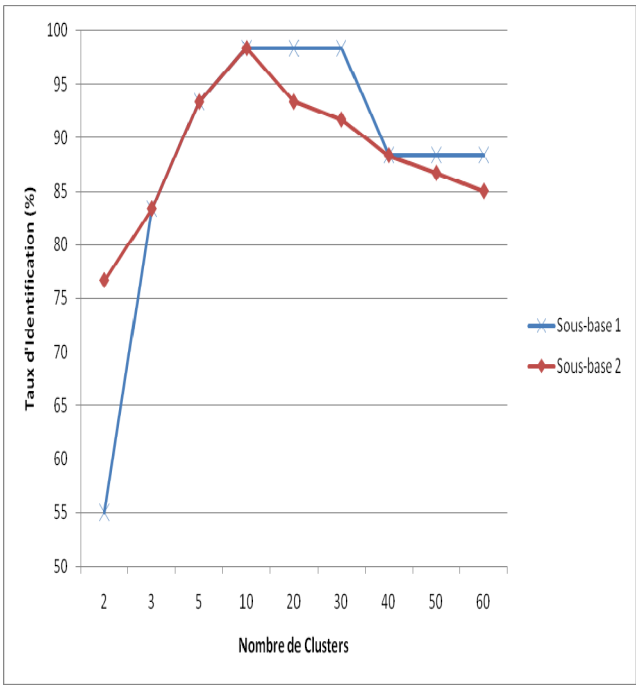


FIG. 7 – Sensibilité de la méthode au nombre de prototypes par lettre.

5 Conclusion

Nous avons proposé une méthode d'identification du scripteur pour des documents contenant de l'écriture en-ligne. Cette méthode donne de très bons résultats sur les bases utilisées. La taille de la base de référence est de 120 scripteurs parmi lesquels, il faut retrouver le scripteur à identifier. La base de test comporte également 120 documents ; parmi ceux-ci un seul scripteur n'est pas retrouvé en première position ; il est toutefois en seconde position. Ces conditions d'évaluation correspondent à un usage réaliste en situation d'entreprise, où de nombreux collaborateurs (enquêteurs, journalistes, secrétaires, …) contribuent à alimenter le système d'information de la société.

La méthode proposée repose sur la spécificité de l'usage des styles allographiques entre les scripteurs. Ces allographes sont examinés au niveau caractère, pour cela la première étape de la méthode consiste à segmenter et reconnaître le texte à ce niveau caractère. A cet effet, un outil industriel automatique a été mis en œuvre, il donne des performances tout à fait suffisantes pour cette tâche. Les études expérimentales menées ont permis de répondre à plusieurs questions, concernant notamment le choix de la métrique dans l'espace de représentation des styles allographiques et aussi du nombre de prototypes nécessaires dans cet espace. Ainsi, nous avons obtenu les meilleurs résultats avec la distance du Chi^2 en utilisant un nombre de 10 prototypes par lettre de l'alphabet en ne considérant que les lettres minuscules. Le passage d'une quantification par un symbole discret correspondant au plus proche prototype d'un caractère à une représentation continue indiquant les probabilités a posteriori que ce caractère soit l'un des $N=10$ prototypes disponibles a permis d'améliorer de façon notable les résultats d'identification. Cela a fait passer le taux d'identification du scripteur de 96,7 % à 99,2 %.

En termes de perspectives, il serait intéressant d'évaluer la robustesse de la méthode vis-à-vis d'une augmentation de la taille de la base des scripteurs de référence. Le nombre de 120 correspond déjà à un nombre significatif, par exemple celui d'un amphi d'étudiants composant à un examen. Bien entendu, en augmentant cette taille, on augmente les risques de confusion inter-scripteurs. Dès lors, pour conserver le niveau élevé de performances actuelles, on peut envisager d'avoir une approche plus ciblée au niveau d'un scripteur. Ici, toutes les lettres minuscules ont été prises en compte et les coefficients de pondération des termes (*idf*) sont communs à l'ensemble des scripteurs. Il serait possible d'adapter dynamiquement l'alphabet utilisé, le nombre de prototypes par lettre et les poids de ces prototypes en fonction du scripteur.

Remerciements

Cette recherche est soutenue conjointement par Nanyang Technological University de Singapore, le programme du Ministère des Affaires Etrangères Merlion PhD et le projet ANR Technologie Logicielle (CIEL 06-TLOG-009).

Bibliographie

[Oviatt 00] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson and D. Ferro, "Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions", *Human-Computer Interaction*, 2000, Vol. 15, No. 4, pp. 263-322.

[Jain 03] A.K. Jain and A. M. Namboodiri, "Indexing and Retrieval of On-line Handwritten Documents", *Proceedings of the 7th International Conference on Document Analysis & Recognition,* 2003, pp.655-659.

[Hochberg 97] J. Hochberg, P. Kelly, T. Thomas and L. Kerns, "Automatic script identification from document images using cluster-based templates", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.19 no.2, Feb 1997, pp.176-181.

[Busch 05] A. Busch, W.W. Boles and S. Sridharan, "Texture for Script Identification", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no.11 Nov 2005, pp. 1720-1732.

[Bulacu 07] M. Bulacu and L. Schomaker, "Text-Independent Writer Identification and Verification Using Textural and Allographic Features", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no.4, Apr 2007, pp. 701-717.

[Niels 07] R. Niels and L. Vuurpijl, Automatic Allograph Matching in Forensic Writer Identification, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 21, No. 1 (2007), pp. 61–81.

[Chan 08] S.K Chan, C. Viard-Gaudin and Y.H Tay "Online Text Independent Writer Identification Using Character Prototypes Distribution", *Proc. of SPIE-IS&T Electronic Imaging: Document Recognition and Retrieval XV*, 2008, vol. 6815, pp.1-9

[Pitak 04] T. Pitak and T. Matsuura, "On-line Writer Recognition for Thai Based on Velocity of Bary center of Pen-point Movement", *Proceedings of IEEE International Conference on Image Processing*, October 2004,pp.889-892.

[Yasushi 03] Y. Yasushi, T. Nagao and N. Komatsu, "Text-indicated Writer Verification Using Hidden Markov Models", *Proceedings of the 7th International Conference on Document Analysis & Recognition*, 2003, pp.329-332.

[He 08] Z. He, X. You and Y. Y. Tang. "Writer identification of Chinese handwriting documents using hidden Markov tree model", *Pattern Recognition* 41, 2008, pp. 1295 – 1307.

[Bensefia 05] A.Bensefia, T.Paquet, L.Heutte, "Handwritten Document Analysis for Automatic Writer Recognition", *Electronic Letters on Computer Vision and Image Analysis*, 2005, vol. 5, no. 2, pp. 72-86.

[Schomaker 04] L. Schomaker and M. Bulacu, "Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 6, June 2004, pp. 787-798.

[Viard-Gaudin 99] C. Viard-Gaudin, P-M Lallican, S. Knerr and P. Binter, "The IRESTE On/Off (IRONOFF) Dual Handwriting Database", *Proceedings of the 5th International Conference on Document Analysis & Recognition*, Sep1999,pp.455-458.

[Vision 07] Vision Objects Industrial Text Recogniser SDK, "MyScript Builder Help", SDK documentation, *http://www.visionobjects.com/about-us/download-center/_263/myscript-products-datasheets.html*, 2007.

[Han 06] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Elsevier, 2006, pp.383-460.

[Hoppner 99] F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, "*Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*", Wiley, 1999, pp. 5-31.

[Salton 88] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information Processing & Management* 24(5), 1988, pp. 513–523.

[Cover 91] T. Cover and J. Thomas, "*Elements of Information Theory"* Wiley, 1991, pp.13-41.